DLP Technologies: Challenges and Future Directions

Nikita Nikitinsky, Tamara Sokolova and Ekaterina Pshehotskaya
InfoWatch
Nikita.Nikitinsky@infowatch.com, Tamara.Sokolova@infowatch.com,
Ekaterina.Pshehotskaya@infowatch.com

ABSTRACT

Currently, many vendors offer multiple data loss prevention solutions, but they need to cope with the challenges of the modern world like access control problem or social network analysis problem to better protect the increasing volume of data. To handle this task DLP-systems have to apply modern statistical algorithms to data protection. In this paper, we will provide an overview of contemporary machine learning algorithms that can help DLP-systems to better detect and analyze data.

KEYWORDS

Machine learning, algorithms, DLP-technologies, clustering, social network analysis, data classification, DLP challenges

1 INTRODUCTION

Nowadays more and more companies face problems of data leakage. Important data (such as confidential or proprietary information) may leak either by improper handling or intentional loss to competitors.
For example, according to Cisco, primary causes of data loss are the following:
"1. Unauthorized application use - almost 70 percent of IT professionals believe the use of unauthorized programs resulted in as many as half of their companies' data loss incidents.
2. Misuse of corporate computers - 44 percent of employees share work devices with others without supervision.
3. Unauthorized physical and network access - 39 percent of IT professionals said they have dealt with an employee accessing unauthorized parts of a company's network or facility.
4. Remote worker security - 46 percent of employees admitted to transferring files between work and personal computers when working from home.
5. Misuse of passwords - 18 percent of employees share passwords with co-workers. That rate jumps to 25 percent in China, India, and Italy." [1]

In order to prevent such cases of data loss different DLP (data loss prevention) systems have been created. Various DLP products may be purchased from multiple vendors like Symantec, InfoWatch, Trend Micro and others. Despite this fact, data leakage prevention problem didn't rivet much attention of academic research community. In this paper, we will discuss current state of DLP technologies, current challenges and future directions of DLP technologies from an algorithmic point of view.

2 CURRENT STATE OF DLP TECHNOLOGIES

Originally DLP systems could detect and analyze only textual data in three ways: regular expressions, keywords, and hashing. Regular expressions were used primarily to recognize data by type, e.g., social security numbers, telephone numbers, addresses, and other data that had a significant amount of structure. Keyword matching was appropriate when a small number of known keywords could identify private data. For example, medical or financial records might meet this criterion. For less structured data, DLP products used hash fingerprinting. The DLP system took as input a set of confidential documents and computed database of hashes of substrings of those documents. The system considered a new document to be confidential if it contained a substring with a matching hash in the database.
Regular expressions were good for detecting well-structured data, but keyword lists could be difficult to maintain and fingerprint-based methods could miss confidential information if it was reformatted or rephrased for different contexts such as email or social networks [2].

There are, however, some attempts to improve the situation, like the usage of machine learning-based tools for image analysis for better fingerprint or photographed credit card detection, but the overall situation in an area of DLP is not so good. [3][4].

3 CURRENT CHALLENGES

The current challenges to be solved by contemporary DLP systems are the following:

1. Encryption challenge. It may be difficult enough to detect and intercept some encrypted confidential data. While encryption provides means to ensure the confidentiality, authenticity and integrity of the data, it also makes it difficult to identify the data leaks occurring over encrypted channels.

2. Access control challenge. It is sometimes not easy to configure and control employees' access to corporate data repositories. That is why, for example, a programmer may access a project that he/she is not involved in to steal some data if an access control system grants full access to all code repositories for all programmers. This challenge is close to the next one.

3. Social network challenge — simple access control approach will not be able to control data flows in a company. This approach is not sufficient to capture heterogeneous communication groups where people can belong to more than one group and, even more, when new communication groups are formed and old ones disappear. That is why it may be impossible to reveal a person leaking data (an «outsider») in a communication or to detect persons having access to limited-access data.

4. New data and customization challenge — it is sometimes not easy to customize a DLP system for a particular customer if the system utilizes old methods of data protection like regular expressions, keywords or digital fingerprints. A customization process thus requires creation of regular expressions, manual keyword analysis and so on — the process may take long time. Furthermore, this process is destined to be repeated as new types of confidential or proprietary data appear.

Most of these challenges may be solved by introducing modern algorithms to DLP systems.

4 FUTURE DIRECTIONS

4.1 Text clustering

Cluster analysis or clustering is a convenient method for identifying homogenous groups of objects called clusters. Objects in a specific cluster share many characteristics, but are very dissimilar to objects not belonging to that cluster. Cluster analysis may be used to resolve an Access Control challenge and, partially, Social network challenge by grouping textual data into clusters and thus — "flagging", for example, any deviations in usual data usage or labeling some data for further DLP purposes.

Among modern clustering and topic modeling algorithms we can name:

LSI (Latent Semantic Indexing) — is an unsupervised machine learning method, which is mostly used for dimensionality reduction. It is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts.

LDA (Latent Dirichlet Allocation) - is also an unsupervised machine learning method, which is mostly used for object clustering. It is a generative model that allows sets of observations to be explained by unobserved

groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. [5]

4.1.1 Clustering evaluation metrics

To evaluate algorithm performance one may use two types of metrics often utilized for cluster analysis purposes:

1. External evaluation metrics. In external evaluation, clustering results are evaluated based on data that was not used for clustering, such as known class labels and external benchmarks. Such benchmarks consist of a set of pre-classified items, and these sets are often created by human (experts). Thus, the benchmark sets can be thought of as a gold standard for evaluation.

We suggest using the following external measurements:

1.1 Jaccard index - also known as the Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

1.2 V-measure score - is an entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied. V-measure is computed as the harmonic mean of distinct homogeneity and completeness scores, just as precision and recall are commonly combined into F-measure.

1.3 Rand index - is a measure of the similarity between two data clusterings. A form of the Rand index may be defined that is adjusted for the chance grouping of elements; this is the adjusted Rand index. From a mathematical standpoint, Rand index is related to the accuracy, but is applicable even when class labels are not used.

1.4 Mutual information score - a variation of mutual information (which is a measure of the variables' mutual dependence) may be used for comparing clusterings. There are many different variants of mutual information metrics (for example, an adjusted variant of the Mutual information score called Adjusted Mutual information score, which is defined in analogy to the adjusted Rand index of two different partitions of a set)

2. Internal Evaluation Metrics. In internal evaluation clustering result is evaluated based on the data that was clustered itself. These methods usually assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters.

We suggest using the following internal measurements:

Silhouette Coefficient — is a measure of how appropriately the data has been clustered and how well each object lies within its cluster. [5]

Dunn index - it is an internal evaluation scheme, where the result is based on the clustered data itself. As do all other such indices, the aim is to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart, as compared to the within cluster variance. For a given assignment of clusters, a higher Dunn index indicates better clustering. [6]

The Davies–Bouldin index. This is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset. [7]

There are some work conducted on enhancing text clustering quality by introducing hybrid algorithms like

LDA+GS+GMM or distributed algorithms to process larger datasets, like Approximate Distributed Hierarchical Dirichlet Processes (AD-HDP), so a DLP system may embed a more complex algorithm than we discussed here therefore improving performance of clustering. [5][8][9]

4.2 Social Network Analysis

Social network analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships.
Applying social network analysis to data leak prevention involves monitoring online communication (email, document and code repositories) to discover the social networks. The discovered social networks are vital in identifying collaborators such as a team of developers working on the same code repository or a group of employees exchanging emails to perform a task. Social network analysis has the potential to discover the types of communication that are not documented as a part of company policy or access control. Social networks may be easily visualized for manual or automatic validation.

To evaluate Social Network for DLP purposes, we can apply the following metrics as the most significant:

4.2.1 Centrality metrics

Centrality metrics discover the most important nodes in a social network:

Degree centrality. Social network researchers measure network activity for a node by using the concept of degrees -- the number of direct connections a node has. [10]

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. It was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network [11]

Eigenvector centrality is a measure of the influence of a node in a network. [10]

4.2.2 Segmentation metrics

Segmentation metrics discover groups in a social network

Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. [12]

Structural cohesion is defined as the minimal number of actors in a social network that need to be removed to disconnect the group. It is thus identical to the question of the node connectivity of a given graph. [13]

4.3 Machine learning algorithms for classification

In the modern world a company may obtain new confidential or proprietary data frequently. Current DLP systems experience problems in detecting such data — technologies like regular expressions or keywords cannot be customized to protect a large flow of diverse information.

That's why new types of algorithms have emerged that enables organizations to use software that learns to detect the types of confidential data that require protection. Through training, this approach will continuously improve the accuracy and reliability of finding protected information.

For DLP purposes, it is vital to classify information into at least two classes (for example, confidential and non-confidential data) and most companies have sample data that may be used to train machine learning algorithms to classify and detect different

types of data. That is why we will discuss supervised learning classification algorithms.

4.3.1 Classification algorithms

The most relevant classification machine learning algorithms for DLP purposes are the following:

Support Vector Machines (SVM) – "are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on." For DLP purposes, we will mostly need Support Vector Classification (SVC) models of SVM. [14]

Decision trees of modern versions like C5.0 or CART (Classification and Regression Trees) - a schematic tree-shaped diagram used to determine a course of action or show a statistical probability. Each branch of the decision tree represents a possible decision or occurrence. The tree structure shows how one choice leads to the next, and the use of branches indicates that each option is mutually exclusive. The structure allows users to take a problem with multiple possible solutions and display it in a simple, easy-to-understand format that shows the relationship between different events or decisions. Modern versions of Decision trees use less memory and are more accurate than older versions like ID3 or C4.5. Although the decision tree approach may be less accurate than the SVM approach, it definitely is more interpretable by human experts. [15][16]

4.3.2 Ensemble and multiclass classification algorithms

The overall quality of the mentioned above algorithms may be significantly increased by using ensemble methods and multiclass classification techniques:

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms [17]

The most popular ensemble methods are based on randomized decision trees. We suggest using a Random forest classification method may suit better for DLP purposes. However, other ensemble learning algorithms are also applicable, for example, Gradient Boosted Regression Trees (GBRT) or Extremely Randomized Trees approach.

Random forest classification - method for classification that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. [18]

Multiclass classification (multinomial classification) is the problem of classifying instances into more than two classes. While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies. There are some techniques to solve the problem:

One-vs-all (One-vs-the-rest) - The strategy consists in fitting one classifier per class. For each classifier, the class is fitted against all the other classes. In addition to its computational efficiency, one advantage of this approach is its interpretability. Since each class is represented by one and one classifier only, it is possible to gain knowledge about the class by inspecting its corresponding classifier. This is the most commonly used strategy.

One-vs-one method - This strategy consists in fitting one classifier per pair of classes. [19]

5 CONCLUSIONS

In this paper, we discussed the current state of contemporary DLP technologies and challenges which DLP technologies faced. We also suggested some algorithms and techniques, which may be implemented in DLP systems to enhance the quality of data analysis and protection.

Although current DLP solutions can somehow protect a significant part of confidential data, the increasing number of electronic means of sharing information and means of information representation implies the increase in the number of methods of confidential data detection. We see the further development of the DLP industry in applying modern machine learning algorithms to data protection. Usage of such algorithms will help us to prevent larger volumes of data.

In this paper, we presented challenges that are being experienced by contemporary DLP systems and ways to solve them:

Cluster analysis algorithms can handle a task of grouping data into clusters for further analysis to deal with the Access Control challenge and help Social Network Analysis algorithms to successfully cope with Social network challenge. Other machine learning classification algorithms as SVM, Decision trees or Random Forests can learn to classify data and thus protect information more efficiently.

REFERENCES

[1] Cisco Systems, Inc., "Data Leakage Worldwide: Common Risks and Mistakes Employees Make"(white paper), 2008

[2] Hart, M., Manadhata, P., and Johnson, R., "Text Classification for Data Loss Prevention, HP Laboratories". HPL-2011-114.

[3] Pshehotskaya, E., Sokolova, T., Ryabov, S., ―New Approaches to Data Classification in DLP Systems―, The International Conference on Computing Technology and Information Management (ICCTIM2014), 209-214, 2014

[4] Symantec, "Machine Learning Sets New Standard for Data Loss Prevention: Describe, Fingerprint, Learn" (white paper: data loss prevention), http://eval.symantec.com/mktginfo/enterprise/white_papers/b-dlp_machine_learning.WP_en-us.pdf

[5] Nikitinsky, N., Sokolova, T., Pshehotskaya, ―E.,Composite Heuristic Algorithm for Clustering Text Data Sets ‖, International Journal of Cyber-Security and Digital Forensics (IJCSDF) 3(3): 153 – 162, 2014

[6] Dunn, J. C., "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". Journal of Cybernetics 3 (3): 32–57, 1973

[7] Davies, D. L.; Bouldin, D. W., "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227, 1979

[8] Newman, D., Asuncion, A., Smyth, P., Welling, M., ―Distributed Algorithms for Topic Models‖, Journal of Machine Learning Research 10, pp. 1801-1828, 2009

[9] Nelson, C., Pottenger, W. M., Keiler, H., and Grinberg, N., "Nuclear Detection Using Higher-Order Topic Modeling.", IEEE International Conference on Technologies for Homeland Security. Waltham, MA. 13-15, 2012.

[10] Opsahl, T., Agneessens, F., Skvoretz, J., "Node centrality in weighted networks: Generalizing degree and shortest paths". Social Networks 32 (3): 245, 2010

[11] Wasserman, S., Faust, K., "Social Networks Analysis: Methods and Applications". Cambridge: Cambridge University Press, 1994

[12] Hanneman, R. A., Riddle, M., "Concepts and Measures for Basic Network Analysis". The Sage Handbook of Social Network Analysis. SAGE. pp. 346–347, 2011

[13] Moody, J., and Douglas R. W., "Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups." American Sociological Review 68(1):103–127, 2003

[14] Cortes, C.; Vapnik, V., "Support-vector networks". Machine Learning 20 (3): 273, 1995

[15] Rokach, L., Maimon, O., "Data mining with decision trees: theory and applications". World Scientific Pub Co Inc, 2008

[16] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., "Classification and regression trees". Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984

[17] Opitz, D.; Maclin, R., "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research 11: 169–198, 1999

[18] Breiman, L., "Random Forests". Machine Learning 45 (1): 5–32, 2001

[19] Bishop, C.M. "Pattern Recognition and Machine Learning". Springer, 2006